A Brief Review of Probability Theory

Probability theory

- Probability theory is incorporated into machine learning, particularly the subset of artificial intelligence concerned with predicting outcomes and making decisions.
- In computer science, functions are used to limit the functions outcome to a value between 0 and 1.
- It is the core concept as well as a primary prerequisite to understanding the ML models and their applications.

- Probability theory is a mathematical framework for quantifying our uncertainty about the world.
- It allows us (and our software) to reason effectively in situations where being certain is impossible.
- Probability theory is the foundation of many machine learning algorithms.

Mathematics of Probability

- Probability is all about the possibility of various outcomes. The set of all possible outcomes is called the sample space.
- The sample space for a coin flip is {heads, tails}. The sample space for the temperature of water is all values between the freezing and boiling point.
- Only one outcome in the sample space is possible at a time, and the sample space must contain all possible values.
- The sample space is often depicted as Ω (capital omega) and a specific outcome as ω (lowercase omega).
- We represent the probability of an event ω as P(ω).

- The probability of any event has to be between 0 (impossible) and 1 (certain), and the sum of the probabilities of all events should be 1.
- The two basic axioms of probability are

$$0 \leq P(\omega) \leq 1$$

$$\sum_{\omega} P(\omega) = 1$$

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

Probability of B

Bayes' Rule

- Bayes' Rule: The chain rule for two variables in two equivalent ways:
- $P(x, y) = P(x|y) \cdot P(y)$
- $P(x, y) = P(y|x) \cdot P(x)$
- In probability theory, the chain rule (also called the general product rule) **permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities.**

- If we set both right sides equal to each other and divide by P(y), we get
- Bayes' rule:

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

- **Expectation:** The expected value, or expectation, of a function h(x) on a random variable x ~ P(x) is the average value of h(x) weighted by P(x). For a discrete x, we write this as

$$\mathbb{E}[h(x)] = \sum_x P(x) \cdot h(x)$$

Variance and Covariance:

- variance is a measure of how much random values vary from their mean.
- Similarly, for functions of random variables, the variance is a measure of the variability of the function's output from its expected value.

$$\mathrm{Var}(h(x)) = \mathbb{E}[(h(x) - \mathbb{E}[h(x)])^2]$$

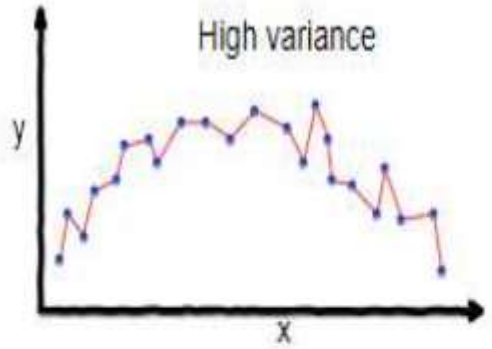
- Machine Learning is an interdisciplinary field that uses statistics, probability, algorithms to learn from data and provide insights which can be used to build intelligent applications

The bias variance trade off

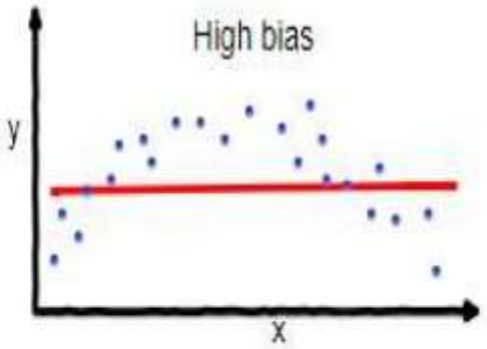
- In machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

- Bias: The bias is known as the difference between the prediction of the values by the ML model and the correct value.

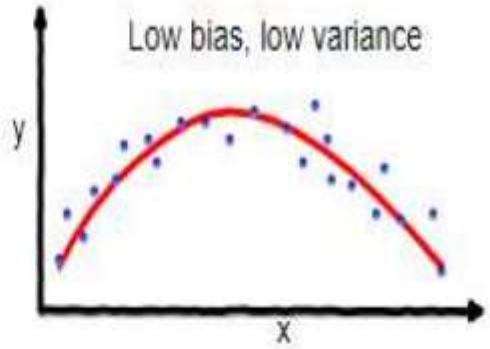
Different types of Biasing



overfitting



underfitting



Good balance

Under fitting of Data(Bias)

- Being high in biasing gives a large error in training as well as testing data.
- Its recommended that an algorithm should always be low biased to avoid the problem of under fitting.
- By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set.
- Such fitting is known as Under fitting of Data.

Overfitting of Data(Variance)

- The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model.
- The model with high variance has a very complex fit to the training data and thus is not able to fit accurately on the data which it hasn't seen before.
- As a result, such models perform very well on training data but has high error rates on test data.
- When a model is high on variance, it is then said to as Overfitting of Data.

Bias Variance Trade off

- While training a data model variance should be kept low
- If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone.
- If algorithms fit too complex (hypothesis with high degree eq.) then it may be on high variance and low bias.

The best fit will be given by hypothesis on the tradeoff point. The error to complexity graph to show trade-off is given as